

---

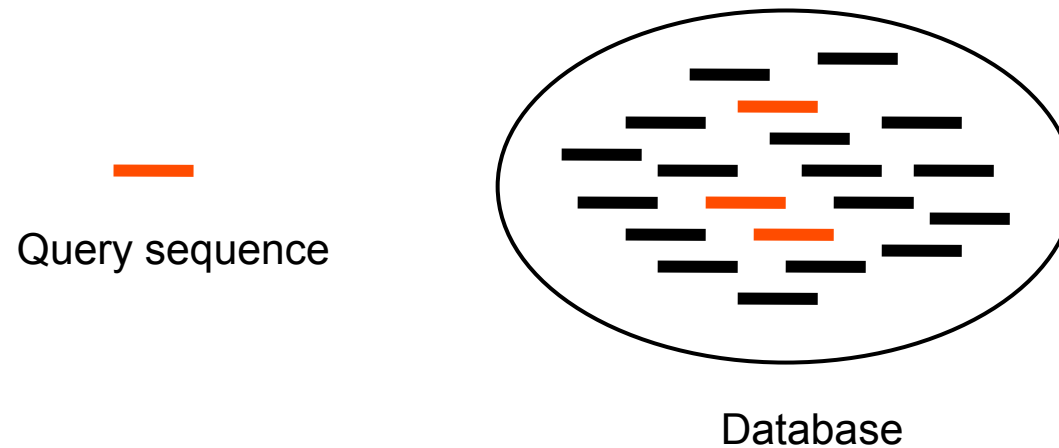
# BLAST

Ulrich Johan Kudahl

# Database searching

---

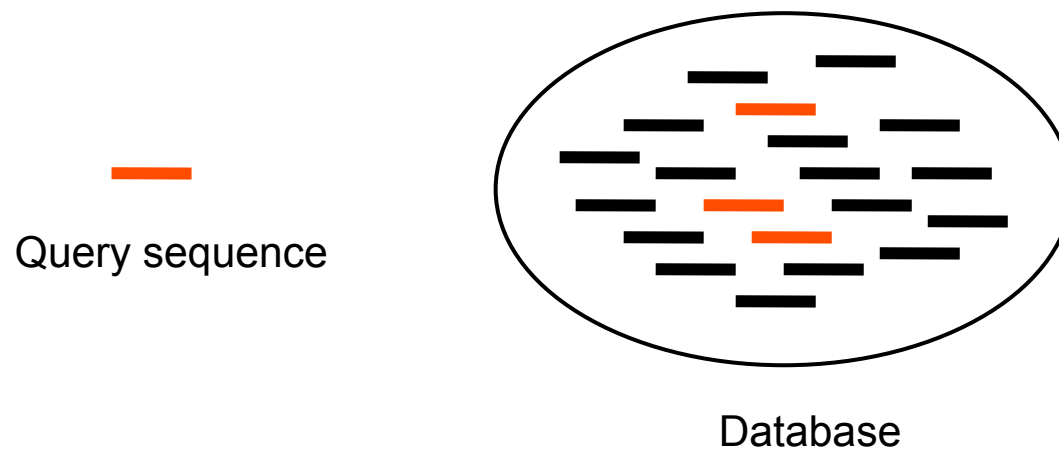
Using pairwise alignments to search  
databases for similar sequences



# Database searching

---

- One Pairwise alignment for each sequence
- Using Local Alignment
- Database size



# Database searching: heuristic search algorithms

---

BLAST (Altschul et al. 1990)

Extremely fast

Two orders of magnitude faster than Smith-Waterman

Uses statistics/likelihood to deal with large database sizes

Very Efficient for finding rough screenings

Uses rapid word lookup methods to completely skip most of the database entries

# BLAST flavors

---

## BLASTN

Nucleotide query sequence

Nucleotide database

Main databases:

- NR (All sequences)
- Human

## BLASTP

Protein query sequence

Protein database

Main databases:

- NR (All sequences)
- Human
- Swissprot

# Searching on the web: BLAST at NCBI

Very fast computer dedicated to running BLAST searches

Many databases that are always up to date (e.g. NR and Human Genome)

Nice simple web interface

The screenshot displays the NCBI BLAST web interface in a browser window. The title bar reads "Protein BLAST: search protein databases using a protein query". The address bar shows the URL "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&". The page header includes the BLAST logo, navigation links (Home, Recent Results, Saved Strategies, Help), and a "My NCBI" link with "Sign In" and "Registered" options. The main content area is titled "Enter Query Sequence" and contains a large text input field for "Enter accession number, gi, or FASTA sequence", a "Clear" button, and a "Query subrange" section with "From" and "To" input fields. Below this is a file upload section with a "Choose File" button and a "Job Title" input field. The "Choose Search Set" section includes a "Database" dropdown menu set to "Non-redundant protein sequences (nr)", an "Organism" input field with a help icon, and an "Entrez Query" input field. The "Program Selection" section features radio buttons for "blastp (protein-protein BLAST)", "PSI-BLAST (Position-Specific Iterated BLAST)", and "PHI-BLAST (Pattern Hit Initiated BLAST)". At the bottom, there is a "BLAST" button, a "Search database nr using Blastp (protein-protein BLAST)" button, and a checkbox for "Show results in a new window". A footer section contains links for "Algorithm parameters", "Copyright", "Disclaimer", "Privacy", "Accessibility", "Contact", and "Send feedback on new interface".

# When is a database hit significant?

---

- Problem:

- Even unrelated sequences can be aligned (yielding a low score)
- How do we know if a database hit is meaningful?
- When is an alignment score sufficiently high?

- Solution:

- Determine the range of alignment scores you would expect to get for random reasons (i.e., when aligning unrelated sequences).
- Compare actual scores to the distribution of random scores.
- Is the real score much higher than you'd expect by chance?

# Distribution of random alignment scores

---

- Software simulation

>P29600|Savinase from Bacillus Lentus

AQSVPWGISRVQAPAAHNRGLTGSQVAVLDTGISTHPDLNIRGGASFVPGEPTQDGN  
GHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMHVA  
NLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGSISYPARYANAMAVGATDQNNNR  
ASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPSWSNVQI  
RNHLKNTATSLGSTNLYGSGLVNAEAATR

>P41363\_Alkaline Protease from Bacillus Halodurans

AAKEPGAQKSAQGINAGNVHKMSGGPIVISVIAIVLVNVVTAPSNIYAVQNMSSASDQ  
TNNTTTASHIEYLKEEYTYTDSIIVGPAIVFANSTEQNVEERQRGAKDIMQGDGVV  
WGGGYAMSNRHMFLTLRSSPSHGVLLLEANISQLGQIKSRNKTCLGGAGGSLN  
ANTRIQVIANLLTNIVRYNGNFGNLNAGTSPALLTLISPINGGSSSINAPASTLAYVK  
WLRGTAVSAEFRIVTVDLTAVVEHARNIGVASEVPYHPQEQHYRVAYLQGPETVNV  
ASAGASDKRNFGFASDVRGIKYVVQALHPLYHISEHLPMDNAAGRSVDEGAAY  
SAIVEYAVSLNELLGEKPANMESKE

[http://www.ebi.ac.uk/Tools/psa/emboss\\_water/](http://www.ebi.ac.uk/Tools/psa/emboss_water/)  
<http://www.cbs.dtu.dk/biotools/SeqShuffle-1.0/>

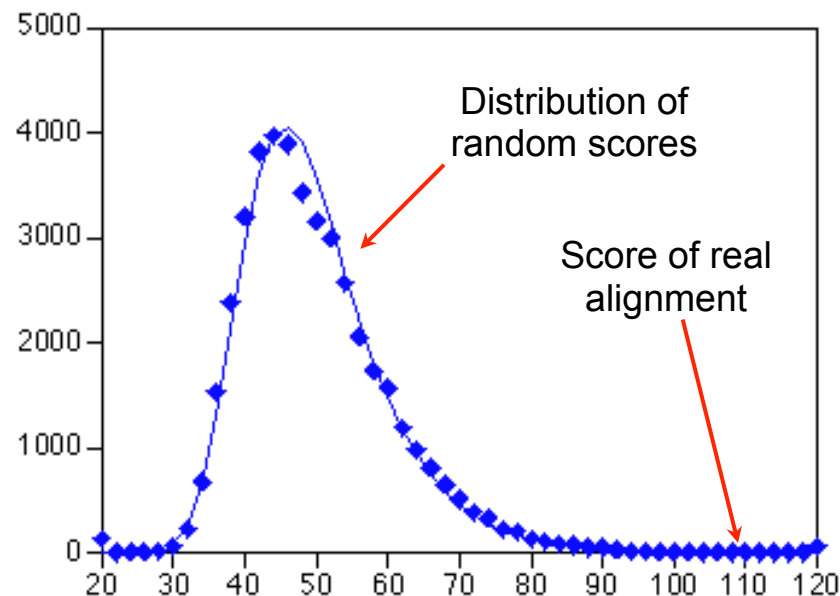


## Significance of alignment score expressed as E-value

---

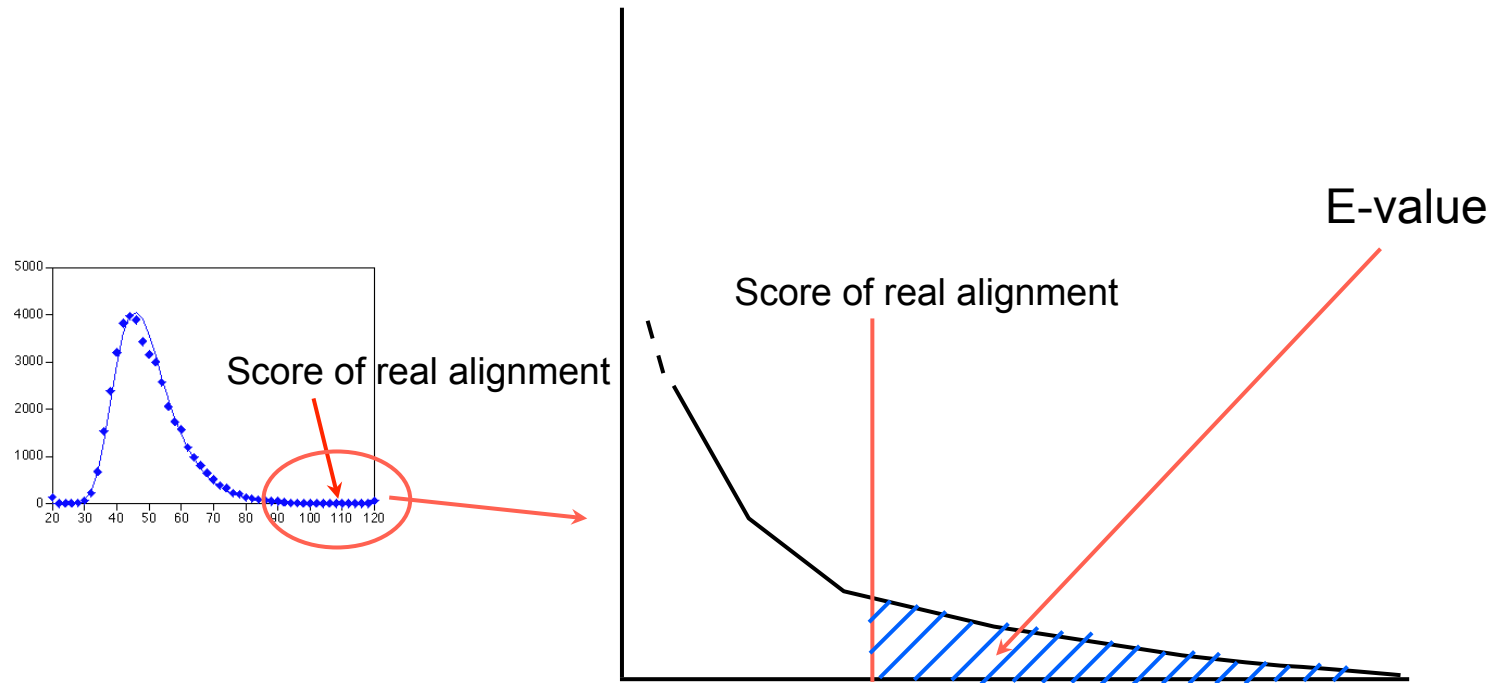
Searching a database of unrelated sequences results in scores following an extreme value distribution

The exact shape and location of the distribution depends on the exact nature of the database and the query sequence



## Significance of alignment score expressed as E-value

---



E-value: the number of random hits with score  $\geq$  real score

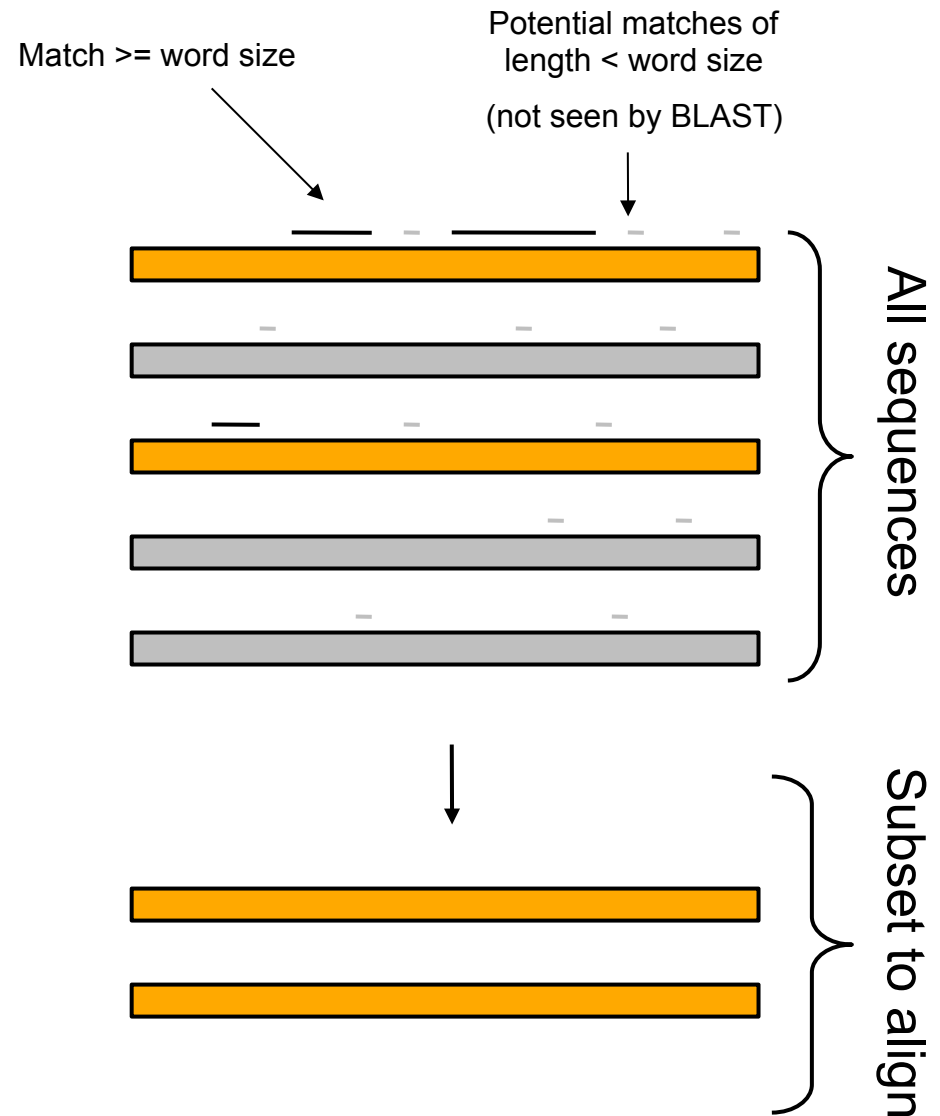
# BLAST heuristics

---

- BLAST speeds up the search  $>100\times$  by pre-screening the database sequences and only performing the full Dynamic Programming on “promising” sequences.
- Promising sequences: database sequences that have substrings (“words”) which also occur in the query sequence (found rapidly using a so-called “suffix-tree”)
- BLASTN and BLASTP use different criteria for overlap required for a sequence to be deemed promising

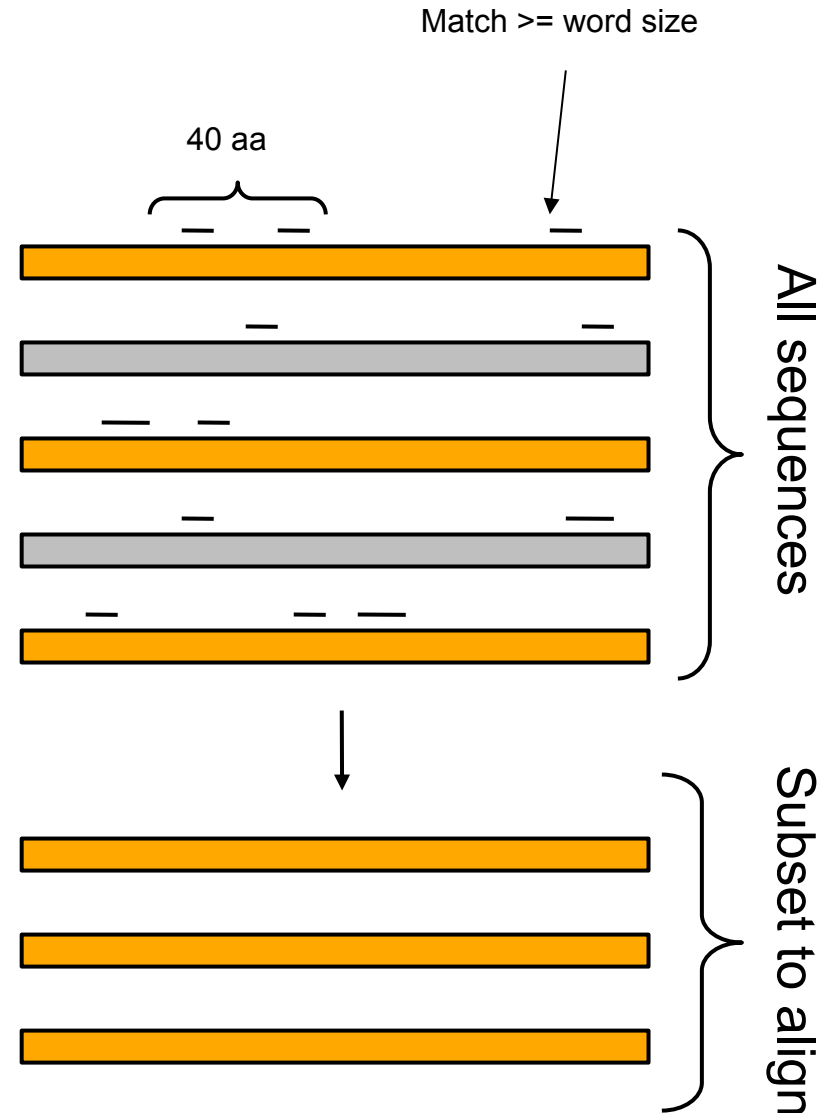
# BLASTN

- Heuristics:
  - Perfectly matching “word” of size  $\geq 11$  bases
- DNA alignment matrix:
  - Match: 1
  - Mismatch: -3



# BLASTP

- Heuristics:
  - 2 x “near match” within a window.
  - Default word length: 3 aa
  - Default window length: 40 aa
- Alignment matrix:
  - Default: BLOSUM 62
- Notice: These alignment matrices incorporate knowledge about protein evolution.



# BLAST

---

- Method for searching for DNA/protein sequences in large databases
- Blast makes a multitude of pairwise alignments and shows the best of them
- The program uses a system to evaluate the chance that the alignment is random with regard to database size and alignment length.